

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION  
International Bureau

## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<b>(51) International Patent Classification 5 :</b> C07H 21/04, C07K 13/00 C12N 15/12, C12P 21/00 C12Q 1/68, G01N 33/574 C07K 15/00	<b>A1</b>	<b>(11) International Publication Number:</b> WO 91/09867  <b>(43) International Publication Date:</b> 11 July 1991 (11.07.91)
<b>(21) International Application Number:</b> PCT/GB90/02020 <b>(22) International Filing Date:</b> 24 December 1990 (24.12.90)  <b>(30) Priority data:</b> 8929097.7 22 December 1989 (22.12.89) GB  <b>(71) Applicant (for all designated States except US):</b> IMPERIAL CANCER RESEARCH TECHNOLOGY LTD [GB/ GB]; Sardinia House, Sardinia Street, London WC2A 3NL (GB).  <b>(72) Inventors; and</b> <b>(75) Inventors/Applicants (for US only) :</b> TAYLOR-PAPADI- MITRIOU, Joyce [GB/GB]; 9 Cedar Road, Berkhamst- ed, Herts HP4 2LA (GB). GENDLER, Sandra [US/ GB]; 20 St James Mansions, West End Lane, West Hampstead, London NW6 2AA (GB). BURCHELL, Joy [GB/GB]; 4 Whites Cottages, Fletching, Uckfield, West Sussex TN22 3SP (GB).		<b>(74) Agents:</b> CRESSWELL, Thomas, Anthony, et al.; J.A. Kemp & Co., 14 South Square, Gray's Inn, London WC1R 5LX (GB).  <b>(81) Designated States:</b> AT (European patent), BE (European patent), CH (European patent), DE (European patent), DK (European patent), ES (European patent), FR (Eu- ropean patent), GB (European patent), GR (European patent), IT (European patent), JP, LU (European pa- tent), NL (European patent), SE (European patent), US.  <b>Published</b> <i>With international search report.</i> <i>Before the expiration of the time limit for amending the</i> <i>claims and to be republished in the event of the receipt of</i> <i>amendments.</i>
<b>(54) Title:</b> MUCIN NUCLEOTIDES  <b>(57) Abstract</b> <p>A nucleic acid fragment comprising a portion of at least 17 contiguous nucleotide bases which portion has a sequence the same as, or homologous to a portion of corresponding length of the sequence of the coding strand as set out in Fig. 1 or the same as, or homologous to a portion of corresponding length of the sequence complementary to the sequence of the coding strand set out in Fig. 1.</p> <pre> -807 tactctctc cccccccc gacggccccc tcaacttcg cggcccgag cgcaggct cccggtacc actaggggc -703 gggagagc cttgacat gttggaggt gcaaggag gacrtagg ttcactgag cccagatta ctcctttag -603 tggaaattc ttcctcact cctccttgc ttctccag gggggacc aggtactag aagtcgac ttgagtgag -553 actgtgggt caggagaga cgggggtgg aacggagac gggagctta gggggggg gcttctcgg gaaatgtgg -453 gggggggg cccaaacta gacactgac cactcatt cagacccct tttctcgg cctctctgt tcatgtgac -403 gggggggg gggggggg agtggggac ctgggggtg gcttccgac ctgtctgac aggtactga cctagctgc -303 ttgtctcc atccccagt ttgtgttgc cctgaggtc aaactagac caggggggc cagttcag actgcccctc -203 ccctctcc cggagccgg ggtgttgg tggagggg aggtcagc gagaacac ggtatgac ggggttga -153 gattagac ctgtacct accagaa ggttggag gaggagag agttatgag taaggagag ggggggtt -53 tgtacatc cactctcgt ctgtcctg gacggggg gggggggg gggggggg atagagcgt aggtcctct -11 gctcgttca ccttcaagc agccagggc tgggtgac ttttctgccc cttccccc ctttctcag ccaactgac 107 ACCGGCGCC CAGTCTGCT TCTTCTGCT GCTGCTGCT ACAGTCTTA CAGTCTGCT CAGTCTGCT CAGTCTGCT 157 tcttctctt aggtctctt cgtctctt ctgtgttt tcttctg cgtctctt cgtctctt cgtctctt 207 tcttctctt aggtctctt cgtctctt ctgtgttt tcttctg cgtctctt cgtctctt cgtctctt 257 tcttctctt aggtctctt cgtctctt ctgtgttt tcttctg cgtctctt cgtctctt cgtctctt 307 tcttctctt aggtctctt cgtctctt ctgtgttt tcttctg cgtctctt cgtctctt cgtctctt 357 tcttctctt aggtctctt cgtctctt ctgtgttt tcttctg cgtctctt cgtctctt cgtctctt 407 tcttctctt aggtctctt cgtctctt ctgtgttt tcttctg cgtctctt cgtctctt cgtctctt 457 tcttctctt aggtctctt cgtctctt ctgtgttt tcttctg cgtctctt cgtctctt cgtctctt 507 tcttctctt aggtctctt cgtctctt ctgtgttt tcttctg cgtctctt cgtctctt cgtctctt 557 tcttctctt aggtctctt cgtctctt ctgtgttt tcttctg cgtctctt cgtctctt cgtctctt 607 tcttctctt aggtctctt cgtctctt ctgtgttt tcttctg cgtctctt cgtctctt cgtctctt 657 tcttctctt aggtctctt cgtctctt ctgtgttt tcttctg cgtctctt cgtctctt cgtctctt 707 ACAGTCTGCT GTCATGCAAG CTCTACCCCA GTGGAGAAA AGGAGCTTC GGCTACCCAG AGAAGTCTGCT TCCGAGCTC 757 TACTGAGAG AATGCTGTA GTATGACCA GAGCTGCTC TCGAGGACA GCGCGCTTC AGGCTGCTC AGGCTGCTC 807 GACAGGATG CAGTCTGCTC CCGGCGCAG ACCAGCTTC AGGCTGCTC GCGAGTCTG GACAGGATG CAGTCTGCTC 857 CGAGTCTGCA GCGAGGCTC GCGGCTGCTC AGGCTGCTC GCGAGTCTG GCGAGTCTG GCGAGTCTG GCGAGTCTG 907 CGAGTCTGCA GCGAGGCTC GCGGCTGCTC AGGCTGCTC GCGAGTCTG GCGAGTCTG GCGAGTCTG GCGAGTCTG 957 CGAGTCTGCA GCGAGGCTC GCGGCTGCTC AGGCTGCTC GCGAGTCTG GCGAGTCTG GCGAGTCTG GCGAGTCTG 1007 CGG </pre>		

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	ES	Spain	MG	Madagascar
AU	Australia	FI	Finland	ML	Mali
BB	Barbados	FR	France	MN	Mongolia
BE	Belgium	GA	Gabon	MR	Mauritania
BF	Burkina Faso	GB	United Kingdom	MW	Malawi
BG	Bulgaria	GN	Guinea	NL	Netherlands
BJ	Benin	GR	Greece	NO	Norway
BR	Brazil	HU	Hungary	PL	Poland
CA	Canada	IT	Italy	RO	Romania
CF	Central African Republic	JP	Japan	SD	Sudan
CG	Congo	KP	Democratic People's Republic of Korea	SE	Sweden
CH	Switzerland	KR	Republic of Korea	SN	Senegal
CI	Côte d'Ivoire	LI	Liechtenstein	SU	Soviet Union
CM	Cameroon	LK	Sri Lanka	TD	Chad
CS	Czechoslovakia	LU	Luxembourg	TG	Togo
DE	Germany	MC	Monaco	US	United States of America
DK	Denmark				

- 1 -

MUCIN NUCLEOTIDES

The present invention relates to nucleotide fragments, polypeptides and antibodies and their use in medical treatment and diagnosis.

5 In International Patent Application no. WO-A-88/05054 there is disclosed a tandem repeat sequence contained in the human polymorphic epithelial mucin (HPEM) gene and nucleotide probes, polypeptides, antibodies and antibody-producing cells which are useful in the diagnosis  
10 and treatment of adenocarcinomas such as breast cancer.

The present inventors have now elucidated the nucleotide base sequence of the gene in the region 5' of the tandem repeat sequence (unless the context implies otherwise, directions such as "5'" or "3'", "upstream" or  
15 "downstream" used herein refer to the non-template strand of the genomic DNA or fragments thereof). The complete sequence of the 1763 nucleotide bases of the non-template strand upstream of and including the first SmaI restriction site in the tandem repeat is set out in Fig.

20 1. The sequence of 1575 nucleotide bases of the non-template strand upstream of and including the first SmaI restriction site in the tandem repeat as set out in Fig. 3 has been extended and some parts have been corrected in the light of repeat experiments. The template strand has  
25 a complementary sequence and it is this strand which is transcribed into RNA during expression of the gene

- 2 -

product.

In addition to conventional transcriptional and translational start sites and intron splicing sites, this sequence contains a number of features which may be  
5 important in the diagnosis and therapy of cancers and in expression of proteins from recombinant vectors. These features will be described below. The amino acid sequence corresponding to the translated portions of this nucleotide sequence gives rise to peptides and thence to  
10 antibodies and antibody-producing cells which may also be useful in such diagnosis and treatment.

In one aspect the present invention provides a nucleic acid fragment comprising a portion of at least 17 contiguous nucleotide bases which portion has a sequence  
15 the same as, or homologous to a portion of corresponding length of the sequence of the coding strand as set out in Fig. 1 or the same as, or homologous to a portion of corresponding length of the sequence complementary to the sequence of the coding strand set out in Fig. 1.

20 As used herein the term "fragment" is intended to include restriction endonuclease-generated nucleic acid molecules and synthetic oligonucleotides.

The nucleic acid fragments of the invention may be single-stranded or double-stranded and they may be RNA or  
25 DNA fragments. Single stranded fragments may be "plus" or coding strands having the sequence of Fig. 1 or a part

- 3 -

thereof or a sequence homologous thereto. Alternatively the single stranded fragments may be "minus" or non-coding strands having a sequence complementary to the sequence of Fig. 1 or a part thereof or a sequence homologous thereto.

5 Double stranded fragments contain a complementary pair of strands, (ie. one plus strand and one minus strand).

RNA fragments according to the invention will, of course, contain uridylic acid ("U") residues in place of the deoxythymidylic acid residues ("T") of the coding  
10 (non-template) strand set out in Fig. 1 or, if complementary to the sequence of the coding strand, they will contain U residues in positions complementary to the adenylic acid ("A") residues in the coding strand set out in Fig. 1.

15 Preferably the nucleic acid fragments of the invention are double-stranded DNA fragments. Single-stranded nucleic acid fragments of the invention are at least 17 nucleotide bases in length. Double-stranded nucleic acid fragments of the invention  
20 are at least 17 nucleotide base pairs in length. Preferably the fragments are at least 20 bases or base pairs in length, more preferably at least 25 bases or base pairs and yet more preferably at least 50 bases or base pairs in length.

25 Statistically it is almost certain that a 17 nucleotide base sequence will be unique so that any

- 4 -

nucleic acid fragment having a contiguous portion of 17 nucleotides of a sequence identical to a portion of corresponding length of the coding strand as set out in Fig. 1, or the same as the non-coding strand complementary to the sequence of Fig. 1, will be new. Fragments according to the invention which are only 17 nucleotides or nucleotide bases in length have a sequence the same as, or complementary to, that set out in Fig.1. Longer fragments of the invention may have a sequence which is homologous to a corresponding portion of the sequence for the coding strand as set out in Fig. 1 or to the complementary non-coding strand.

Preferably nucleic acid fragments according to the invention have at least 75% sequence homology with a corresponding portion of the sequence of Fig. 1 or the complementary non-coding strand, for instance 80 or 85%, more preferably 90 or even 95% homology. Differences may arise through deletions, insertions or substitutions. In addition to containing a portion homologous to or the same as the sequence of the coding strand in Fig. 1 or complementary non-coding strand, the nucleic acid fragments of the invention may include sequences completely unrelated to that in Fig. 1.

Particular features of interest within the coding strand in Fig. 1 are set out in Tables 1 to 3 below:

- 5 -

TABLE 1: Signal Sequences

	Location*	Sequence in PEM	Significance
5	1-2	CG	transcriptional start site
	73-75	ATG	translational start signal
	131-132	GT	start of first intron
	631-632	AG	end of first intron
10	100-130 } and        } 633-637 }	TTCCTGCTGCTGCT- CCTCACAGTGCTTA- CAG...TTGTT	Signal sequence, interrupted by first intron (first intron indicated by "...").
	955-960	CCCGGG	SmaI site at start of tandem repeat

15 Footnotes to Tables 1 and 2

+ In the consensus sequences : R is A or G

N is A, C, G or T

W is A or T

X is

20

Y is C or T

\* Locations are of the 5' base of the indicated PEM  
sequence numbered as in Fig. 1.

- 6 -

TABLE 2 : Regulatory elements within the 5' flanking sequence

Regulatory element	Consensus Sequence <sup>+</sup>	Sequence in PEM	Location*
SP1	GGGCGG	GGGCGG GGGCGG GGGCGG GGGCGGGCGGGCGGG	-727 -397 -94 -54
SV40 enhancer element			
a	ATGTGTGT	CTGTGGGT	-562
b	GCATGCAT	GCCTGCCT	+25
c	GTGGATAG	GTGGAGAG	-702
AP-1	CTGACTCA G    A	GTGACCAC CTGCTTCA GTGCCTAG CTGCCTGA	-739 -418 -61 +27
AP-2	CCCCAGGC G    G	ACCCAGGC CACCGGGC	-597 +77
NF1/CTF	TTGGCTNNNAGCCAA	TTGGCTTTCTCCAA	-618
Glucocorticoid regulatory element:			
Core sequence	TGTTCT	TGTTCT TGTTCC	+38 -321
Consensus sequence	GGTACANNNTGTTCT	GCCTGAATCTGTTCT AGCTGGCTTTGTTCC	+29 -330
CACCC factor	CACCC	CACCC CACCC	+54 +84
Progesterone receptor consensus sequence	ATTCTCTGT	ACTCCTCTCC ACTCCTCCTT ATTCTCTGGC	-802 -626 -432
Estrogen consensus sequence	GGTCANNNTGACC	GCTCCCGGTGACC	-746
RNA Polymerase III Box A	RRYNNARYXGG	GACCTAGCTGG AGTGGAGTGGG	-335 -388
Box B	GWTCRANNC	GTCCAGAC	-260
Enhancer sequences:			
Interferon- $\beta$ seq	GGAAATTCCTCTG	GGAAATTTCTTCC	-642
CMV enhancer	GGAAAGTCCCGTT	GGAAAGTCCGGCT	-585



- 7 -

The sequence in Fig. 1 also includes two sites occurring in the promoter region and in the first intron having 70 to 80% homology with the mammary consensus, sequence (Rosen, J.M. in "The Mammary Gland, Development, Regulation and Function", Ed. Nevill, M.C. and Daniel, C.W. Plenum Press, pp 301-322). These sites are set out in Table 3 below:

TABLE 3

Location	Sequence
10	
-289 to -274	***       * * AGGCTAAACTAGAGC
+230 to +245	*       ** ** GTAAGAATTGCAGACA
15	
Consensus	RGAAGRAAANTGGACA

Positions are numbered in accordance with Fig. 1.

\* indicates a mismatch with the consensus sequence.

In the consensus sequence:- R is A or G.

20

N is A, C, G or T.

Preferred fragments according to the present

- 8 -

invention include the transcriptional and translational start signals, "TATAA" box and at least one of the regulatory elements (transcription factor binding sites) set out in Table 2 above. More preferably these fragments contain 2 or  
5 more, for instance 3, 4 or 5 of the regulatory elements in addition to the TATAA box or even all of the regulatory elements set out in Table 2. Those fragments containing more than one of the regulatory elements of Table 2 preferably also preserve the relative spacings of those sites from one  
10 another and from the TATAA box and transcriptional and translational start signals.

Other preferred fragments of the invention contain at least one of the regions homologous to the mammary consensus sequences as set out in Table 3. Preferably these fragments  
15 contain both of the regions having homology with the mammary consensus sequences as set out in Table 3. Those fragments containing both regions having homology with the mammary consensus sequence preferably also preserve the relative spacing of those regions, as found in Fig. 1, from one  
20 another and from the TATAA box and transcriptional and translational start signals.

Yet further preferred fragments according to the invention comprise the TATAA box, the transcriptional and translational start signals, at least one and preferably two  
25 or more of the regulatory elements as set out in Table 2 and at least one and preferably both of the regions having

- 9 -

homology with the mammary consensus sequence as set out in Table 3. Yet more preferably these fragments also preserve the relative spacing of the features from Tables 1, 2 and 3. Particularly preferred fragments according to the invention

5   comprise the sequence upstream of the TATAA box as set out in Fig. 1 together with, and downstream thereof, transcriptional and translational start signals and a polypeptide coding sequence in correct reading frame register with the promoter sequences and the TATAA box,

10   transcriptional and translational start signals. The coding sequence may encode a part or parts of the polypeptide encoded by the mucin gene, for instance a part or parts thereof other than the tandem repeat sequence, or polypeptides unrelated to that encoded by the mucin gene.

15         Other particularly preferred fragments according to the present invention comprise promoter sequences, a TATAA box, transcriptional and translational start signals and, downstream thereof and in correct reading frame register therewith a coding sequence corresponding to a portion of

20   the mucin gene, for instance corresponding to the first exon (corresponding to bases (1 to 130 of Fig.1.) or a part thereof and/or the second exon (corresponding to bases 633 onwards in Fig.1.) or a part thereof, for instance a part thereof other than the tandem repeat sequence as set out in

25   WO-A-88/05054.

In an especially preferred aspect the fragments

- 10 -

contain (i) the first 26 bases (bases 1 to 26 of Fig. 1) or  
(ii) the whole of the first exon (bases 1 to 130 of Fig.1.)  
and/or (iii) the splicing/ligating sites for the first intron  
set out in Table 1 and a non-coding sequence between these  
5 sites. The non-coding sequence may be the same as or  
different to the sequence of the first intron as shown in  
Fig. 1. Preferably it is the same.

Other preferred fragments of the invention comprise  
at least a portion of the first intron (bases 231 to 632 of  
10 Fig. 1). Further preferred fragments of the invention  
comprise at least a portion of the 5'-flanking sequence  
upstream of base -423 of Fig. 1.

Other preferred fragments of the invention comprise a  
portion of the sequence of Fig. 1 corresponding to a portion  
15 of the sequence of Fig. 3.

Further preferred fragments of the invention comprise  
a combination of any two or more of the foregoing preferred  
features.

Fragments according to the present invention  
20 containing functional coding sequences for a least a part of  
the first or second exons set out in Fig. 1 are useful in the  
production of polypeptides corresponding to a part or all of  
the mucin gene product. Such polypeptides are, in turn  
useful as immunogenic agents for instance in active  
25 immunisation against Human Polymorphic Epithelial Mucin  
(HPEM) for the prophylactic or therapeutic treatment of

- 11 -

cancers or raising antibodies for use in passive immunisation and diagnosis of cancers. For use in such methods the fragment, which codes for a polypeptide chain substantially identical to a portion of the mucin core protein, may be

5 extended at either or both the 5' and 3' ends with further coding or non-coding nucleic acid sequence including regulatory and promoter sequences, marker sequences, and splicing or ligating sites. Coding sequences may code for other portions of the mucin core protein chain (for instance,

10 other than the tandem repeat) or for other polypeptide chains. The fragment according to the invention, together with any necessary or desirable flanking sequences is inserted, in an appropriate open reading frame register, into a suitable vector such as a plasmid, or cosmid or a viral

15 genome (for instance vaccinia virus genome) and is then expressed as a polypeptide product by conventional techniques. In one aspect the polypeptide product may be produced by culturing appropriate cells transformed with a vector, harvested and used as an immunogen to induce active

20 immunity against the mucin core protein [Tartaglia et al., Tibtech, 6, 43: (1988)].

Fragments according to the present invention incorporating regulatory elements of Table 2 and/or mammary consensus sequences of Table 3 may be used in securing

25 tissue-specific expression of functional coding sequences in appropriate reading frame register downstream of the

- 12 -

regulatory elements and/or associated with the mammary consensus sequences. Such fragments may therefore be used to express parts or the whole of the mucin gene or any other coding sequence in cells of epithelial origin. Applications  
5 of this are in therapy and immunisation where such fragments and associated coding sequences are administered to patients such that the coding sequence will be expressed in epithelial tissues leading to a therapeutic effect or an immune reaction by the patient against the polypeptides.

10 The fragments may be presented as inserts in a vector such as viral genomic nucleic acid and introduced into the patients by inoculation of the vector for instance as a modified virus. The vector then directs expression of the polypeptide in vivo and this in turn serves as a therapeutic  
15 agent or as an immunogen to induce active immunity against the polypeptide. This strategy may be adopted, for instance, to secure expression of polypeptides encoded by the HPEM gene for treatment or prophylaxis of adenocarcinomas such as breast cancer or to secure tissue  
20 specific expression of other peptides under control of the regulatory sequences of Table 1, for instance by administration of a modified vaccinia virus containing the fragment and coding sequences in its genomic DNA. RNA fragments of the invention may similarly be used by  
25 administration via a retroviral vector. Selection of tissue specific virus vectors to carry the fragments of the

- 13 -

invention and coding sequences will further restrict expression of the polypeptide to desired target tissues.

Fragments of the invention may also be used to control expression of oncogenic proteins in experimental transgenic animals. Thus, for instance, a transgenic mouse having an oncogene such as ras, erbB-2 or int 2 expressed under control of the present tissue specific fragments may develop breast tumours and be useful in testing diagnostic agents such as tumour localisation and imaging agents and in testing therapeutic agents such as immunotoxins.

Nucleic acid fragments according to the invention are also useful as hybridisation probes for detecting the presence of DNA or RNA of corresponding sequence in a sample. For use as probes fragments are preferably labelled with a detectable label such as a radionuclide, enzyme label, fluorescent label or other conventional directly or indirectly detectable labels. For some applications, the probes may be bound to a solid support. Labelling of the probes may be achieved by conventional methods such as set out in Matthews *et al.*, Anal. Biochem. 169: 1-25 (1988).

In further aspects, the present invention provides cloning vectors and expression vectors containing fragments according to the present invention. The vectors may be, for instance, plasmids, cosmids or viral genomic DNA. The present invention further provides host cells containing

- 14 -

such cloning and expression vectors, for instance epithelial cells transformed with functional expression vectors containing expressible fragments according to the invention.

The invention further provides nucleic acid fragments ,  
5 which encode polypeptides as defined below. Such fragments may be fragments as hereinbefore defined. However, in view of the redundancy of the genetic code, nucleic acid sequences which differ slightly or substantially from the sequence of Fig. 2 may nevertheless encode the same  
10 polypeptide.

The nucleic acid fragments of the invention may be produced de novo by conventional nucleic acid synthesis techniques or obtained from human epithelial cells by conventional methods, Huynh et al., "DNA Cloning: A  
15 Practical Approach" Glover, D.M. (Ed) IRL, Oxford, Vol 1, pp49-78 (1985).

The invention therefore also provides probes, vectors and transformed cells comprising nucleic acid fragments as hereinbefore defined for use in methods of treatment of the  
20 human or animal body by surgery or therapy and in diagnostic methods practiced on the human or animal body and for use in the preparation of medicaments for use in such methods. The invention also provides methods for treatment of the human or animal body by surgery or therapy and diagnostic methods  
25 practiced in vivo as well as ex vivo and in vitro which comprise administering such fragments, probes, vectors or



- 15 -

transformed cells in effective non-toxic amount to a human or other mammal in need thereof.

Processes for producing fragments according to the invention and probes, vectors and transformed cells  
5 containing them and processes for expressing polypeptides encoded by, or under the regulatory control of, fragments of the invention also form aspects of the invention.

The invention further provides a polypeptide comprising a sequence of at least 5 amino acid residues encoded by the  
10 coding portion of the DNA sequence as indicated in Fig. 2. Polypeptides according to the invention preferably have a sequence of at least 10 residues, for instance at least 15, more preferably 20 or more residues and most preferably all the residues shown in Fig. 2.

15 The polypeptide may additionally comprise N-terminal and/or C-terminal sequences not encoded by the DNA sequence indicated by Fig. 2.

Polypeptides of the invention containing more than 5 amino acid residues encoded by the DNA sequence in Fig. 2  
20 may include minor variations by way of substitution, deletion or insertion of individual amino acid residues. Preferably such polypeptides differ at not more than 20% preferably not more than 10% and most preferably not more than 5% of residues in a contiguous portion corresponding to  
25 a portion of the sequence in Fig. 2.

The invention further provides polypeptides as

- 16 -

defined above modified by addition of a linkage sugar such as N-acetyl galactosamine on serine and/or threonine residues and polypeptides modified by addition of oligosaccharide moieties to N-acetyl galactosamine or via other linkage sugars. Optionally modified polypeptides linked to carrier proteins such as keyhole limpet haemocyanin, albumen or thyroglobulin are also within the invention.

Polypeptides according to the invention may be produced de novo by synthetic methods or by expression of the appropriate DNA fragments described above by recombinant DNA techniques and expressed without glycosylation in human or non-human cells. Alternatively they may be obtained by deglycosylating native human mucin glycoprotein (which itself may be produced by isolation from samples of human tissue or body fluids or by expression and full processing in a human cell line) [Burchell et al., Cancer Research, 47: 5467-5482, (1987), Gendler et al., P.N.A.S., 84: 6060-6064, (1987)], and digesting the core protein. The polypeptides of the invention are useful in active immunisation of humans, for raising antibodies in animals for use in passive immunisation, diagnostic tests, tumour localisation and, when used in conjunction with a cytotoxic agent, for tumour therapy.

The invention further provides antibodies against any of the polypeptides described above.

- 17 -

As used hereafter the term "antibody" is intended to include polyclonal and monoclonal antibodies and fragments of antibodies bearing antigen binding sites such as the F(ab')<sub>2</sub> fragments as well as such antibodies or fragments thereof which have been modified chemically or genetically in order to vary the amino acid residue sequence of one or more polypeptide chains, to change the species specific and/or isotype specific regions and/or to combine polypeptide chains from different sources. Especially in therapeutic applications it may be appropriate to modify the antibody by coupling the Fab, or complementarity-determining region thereof, to the Fc, or whole framework, region of antibodies derived from the species to be treated (e.g. such that the Fab region of mouse monoclonal antibodies may be administered with a human Fc region to reduce immune response by a human patient) or in order to vary the isotype of the antibody (see EP-A-0 239 400). Such antibodies may be obtained by conventional methods [Williams, Tibtech, 6:36, (1988)] and are useful in diagnostic and therapeutic applications, such as passive immunisation.

The term "antibodies" used herein is further intended to encompass antibody molecules or fragments thereof as defined above produced by recombinant DNA techniques as well as so-called "single domain antibodies" or "dAbs" such as are described by Ward, E.S. et al., Nature, 341:544-546

- 18 -

(1989) which are produced in recombinant microorganisms, such as Escherichia coli, harboring expressible DNA sequences derived from the DNA encoding the variable domain of an immunoglobulin heavy chain by random mutation introduced, for instance, during polymerase chain reaction amplification of the original DNA. Such dAbs may be produced by screening a library of such randomly mutated DNA sequences and selecting those which enable expression of polypeptides capable of specifically binding the polypeptides of the invention or HPEM core protein.

Antibodies according to the present invention react with HPEM core protein, especially as expressed by colon, lung, ovary and particularly breast carcinomas, but have reduced or no reaction with corresponding fully processed HPEM. In a particular aspect the antibodies react with HPEM core protein but not with fully processed HPEM glycoprotein as produced by the normal lactating human mammary gland.

Antibodies according to the present invention preferably have no significant reaction with the mucin glycoproteins produced by pregnant or lactating mammary epithelial tissues but react with the mucin proteins expressed by mammary epithelial adenocarcinoma cells. These antibodies show a much reduced reaction with benign breast tumours and are therefore useful in diagnosis and localisation of breast cancer as well as in therapeutic methods.

- 19 -

Further uses of the antibodies include diagnostic tests of assays for detecting and/or assessing the severity of breast, colon, ovary and lung cancers.

The antibodies may be used for other purposes  
5 including screening cell cultures for the polypeptide expression product of the human mammary epithelial mucin gene, or fragments thereof, particularly the nascent expression product. In this case the antibodies may conveniently be polyclonal or monoclonal antibodies.

10 The invention further provides antibodies linked to therapeutically or diagnostically effective ligands. For therapeutic use of the antibodies the ligands are lethal agents to be delivered to cancerous breast or other tissue in order to incapacitate or kill transformed cells. Lethal  
15 agents include toxins, radioisotopes and "direct killing agents" such as components of complement as well as cytotoxic or other drugs.

For diagnostic applications the antibodies may be linked to ligands such as solid supports and detectable  
20 labels such as enzyme labels, chromophores, fluorophores and radioisotopes and other directly or indirectly detectable labels. Preferably monoclonal antibodies are used in diagnosis.

Antibodies according to the present invention may be  
25 produced by inoculation of suitable animals with a polypeptide as hereinbefore described. Monoclonal

- 20 -

antibodies are produced by known methods, for instance by the method of Kohler & Milstein [Nature, 256: 495-497 (1975)] by immortalising spleen cells from an animal inoculated with the mucin core protein or a fragment thereof, usually by fusion with an immortal cell line (preferably a myeloma cell line), of the same or a different species as the inoculated animal, followed by the appropriate cloning and screening steps.

Antibody-producing cells obtained from animals inoculated with polypeptides of the invention and immortalised such cells form further aspects of the invention.

The invention further provides polypeptides, antibodies and antibody producing cells, such as hybridomas, as hereinbefore defined for use in methods of surgery, therapy or diagnosis practiced on the human or animal body or for use in the production of medicaments for use in such methods. The invention also provides a method of treatment or diagnosis which comprises administering an effective non-toxic amount of a polypeptide or antibody as hereinbefore described to a human or animal in need thereof.

Processes for producing polypeptides according to the invention whether by expression of nucleic acid fragments of the invention or otherwise, and for producing antibodies or fragments thereof and for producing antibody-producing cells such as immortalised cells, form further aspects of the

- 21 -

invention.

The invention further provides a diagnostic test or assay method comprising contacting a sample suspected to contain abnormal human mucin glycoproteins with an antibody  
5 as defined above. Such methods include tumour localisation involving administration to the patient of the antibody bearing detectable label or administration of an antibody and, separately, simultaneously or sequentially in either order, administering a labelling entity capable of  
10 selectively binding the antibody or fragment thereof. Diagnostic test kits are provided for use in diagnostic tests or assays and comprise antibody and, optionally, suitable labels and other reagents and, especially for use in competitive assays, standard sera.

15 The invention will now be illustrated with reference to the figures of the accompanying drawings in which:

Fig. 1. shows the deoxynucleotide base sequence of the 1763 bases upstream of and including the first SmaI restriction  
20 site in the tandem repeat sequence of WO-A-88/05054 using the conventional symbols A, C, G and T for the bases of the non-template strand. The base sequence is arranged in blocks of ten. Untranscribed sequence is in lower case, transcribed sequence is in upper case. The SP1 regulatory  
25 elements (Table 2), TATAA box, transcriptional and translational start sites (Table 1) are underlined.

- 22 -

Fig. 2. shows the sequence of the non-template strand commencing from the transcriptional start site, (residue 1 in Fig. 1.) and excluding the sequence of the first intron (bases 131 to 632 of the sequence in Fig.1.). Fig.2 also shows the predicted sequence of the polypeptide using the conventional 1 letter symbols for the amino acid residues. Amino acid residues are numbered down the left-hand side and nucleotide bases down the right hand side. The signal sequence is underlined. The sequences end at the first SmaI site in the tandem repeat.

Fig. 3. shows the deoxy nucleotide base sequence of the 1575 bases upstream of and including the first SmaI restriction site in the tandem repeat sequence of WO-A-88/05054 using the conventional symbols A, C, G and T for the bases of the non-template strand. The base sequence is arranged in blocks of ten in non-coding regions. The exon sequences are shown in blocks of three and translated codons are underlined. The start positions of exons 1 and 2, intron 1 and the signal sequence for exon splicing are numbered and labelled. Other features mentioned in Tables 1 and 2 are boxed. The sequence finishes with the first SmaI site of the tandem repeat sequence.

The present invention does not extend to fragments,



- 23 -

polypeptides and antibodies or related materials such as vectors and cells, which are specifically disclosed in WO-A-88/05054 or WO-A-90/05142, nor to the CDNA fragment whose sequence is indicated in Abe, M. et al., in Biochemical and  
5 Biophysical Research Communications, 165(2): 644-649 (1989).

The invention will now be illustrated by the following Examples:

EXAMPLE 1

In an attempt to obtain clones with 5' unique  
10 sequences, two gt10 libraries were screened with a probe for the tandem repeat. All the clones obtained lacked any non-repetitive sequence at the 5' terminus. Thus, a different strategy was adopted. To obtain 5' sequence we synthesized the cDNA corresponding to the 5' end of breast  
15 cancer cell line transcript using anchored-polymerase chain reaction (A-PCR). The A-PCR procedure [Loh, E.Y. et al., Science, 243: 217-220, (1989)] was used to synthesize cDNA corresponding to the 5' end of the transcript. For the 5' end clones total RNA (5 µg) prepared by the guanidinium  
20 isothiocyanate method [Chirgwin, J.M. et al., Biochem., 18: 5294-5299 (1979)] was used for first strand synthesis using a breast cancer cell line (BT20) transcript with AMV-reverse transcriptase (Life Sciences) in a 40 µl reaction mixture

- 24 -

[Okayama, H. and Berg, P., Mol. Cell. Biol., 2: 161-170 (1982)] containing 1 µg of an oligonucleotide primer made to the tandem repeat (5'CCAAGCTTGGAGCCCGGGGCCGCTGGTGTCCGG3'). The total RNA was subjected to reverse transcription, and the products were precipitated with spermine. A poly(dG) tail was introduced with terminal deoxy-transferase (500 U/ml, Pharmacia). Amplification was performed with *Thermus aquaticus* polymerase (Perkin Elmer Cetus) in 100 µl of the standard buffer supplied. The primers included the tandem repeat primer and for the poly(dG) end, a mixture of the AN polyC primer (5'GCATGCGCGCGGCCGCGGAGGCCCCCCCCCCCC3') and the AN primer (5'GCATGCGCGCGGCCGCGGAGGCC3') at a ratio of 1:9. Following an initial denaturation at 94°C for 5 min, the reaction was annealed at 55°C for 2 min, extended at 72°C for 2.5 min and denatured at 94°C for 1.5 min. Amplification was performed for 30 cycles, and the product was precipitated with ethanol. The DNA was sequentially cut with HindIII and SacII, separated on a 1.2% agarose Gel and the band of approximately 550 bp was purified onto DEAE membrane (Schleicher and Schuell), ligated into pBS-SK<sup>+</sup> and transformed into bacteria XL-1 (Stratagene). This plasmid will be referred to as pBS-5'PEM. All restriction enzymes used were obtained from New England Biolabs Inc., oligonucleotide primers and probes were synthesized on an Applied Biosystems 380B DNA synthesizer.

Four colonies were selected for sequencing, and the

- 25 -

sequences agreed with each other and with sequence obtained from genomic clones of the region. A Leader sequence of 72 bp preceded the first ATG which was in-frame with the reading frame of the tandem repeat as previously determined (Fig. 1), and the sequence preceding first ATG, CCACCATGA, agrees with the Kozak consensus sequence (Kozak, M., Nucl. Acids. Res., 12: 857-872 (1984)).

The primer extension technique was use to map precisely the position of the capsite. A 21 bp oligonucleotide primer (5'AGACTGGGTGCCCCGGTGTCAT3') corresponding to nucleotides 73 to 93 ending at the A of ATG (Fig. 1) was end-labelled with [ $\gamma$ - $^{32}$ P]ATP (> 5000 Ci/mmol, Amersham International plc) using T4 poly-nucleotide kinase (Pharmacia) and precipitated three times with equal volumes of 4 M ammonium acetate to remove free [ $\gamma$ - $^{32}$ P]ATP from the kinased oligonucleotide. Labelled primer ( $1 \times 10^5$  dpm at  $1 \times 10^7$  dpm/pmole) was annealed to 40  $\mu$ g of total BT 20 RNA in 120 mM sodium chloride at 95°C for 5 min, held at 65°C for 1 h and cooled to room temperature. The annealed primer was extended using 18 units of reverse transcriptase in 50mM Tris pH 8.3 at 45°C, 6 mM magnesium acetate, 10 mM dithiothreitol, 1.8 mM dNTPs in a total volume of 50  $\mu$ l at 45°C for 1h. The reaction was stopped by the addition of 50 mM EDTA and the RNA digested by treatment with RNase-A at 400 $\mu$ g/ml for 15 min at 37°C. The samples were than phenol:chloroform extracted prior to ethanol precipitation

- 26 -

and electrophoresed on a standard 6% sequencing gel yielding two bands which mapped to two C's, 72 and 71 bases upstream of the ATG. The sequencing ladder was single-stranded control DNA (M13mp18) from the Sequenase kit (US Biochemical Corp.).

The most prominent product was 72 bp, equal to the number of base pairs from the 5' end of the oligonucleotide primer to the 5' end of the PCR-derived clone, thus confirming that the cDNA represents the entire length of its corresponding cellular mRNA 5' to the tandem repeat. The presence of a second band may be due to interference with reverse transcriptase by methylation of the C at base 71, since it forms a CpG dinucleotide. Under identical conditions, no primer extension product was seen using RNA from Daudi cells which do not express the PEM mucin.

### Cloning

A plasmid library, grown in DH1α cells (RecA-), was used instead of a lambda library, because of the possibility of recombination occurring when lambda is grown in RecA+ cells. This recombination might have been expected, since a part of the tandem repeat sequence (GCTGGGGG) is closely related to the chi sequence (GCTGGTGG) of lambda phage which has been implicated as a hotspot for RecA-mediated recombination in E.coli.

- 27 -

Nucleotide sequence of cDNA clones

Fig 1. shows the DNA sequence from the 5' A-PCR-derived clone, including the consensus sequence of the tandem repeat. Sequences were determined in both  
5 directions. The region of conserved tandem repeats was not sequenced in full, although a cDNA tandem repeat clone obtained previously had been circularised, sonicated and about 40 clones sequences [(Gendler et al., J. Biol. Chem., 263:12820-12823 (1988)].

- 28 -

Predicted amino acid sequence and composition of the PEM  
core protein.

The core protein amino acid composition is dominated  
5 by the amino acid composition of the tandem repeat. Serine,  
threonine, proline, alanine and glycine account for about  
60% of the amino acids.

The deduced sequence of the PEM core protein consists  
of distinct regions including (1) the N-terminal region  
10 containing a hydrophobic signal sequence and degenerate  
tandem repeats and (2) the tandem repeat region itself.  
At the N-terminus a putative signal peptide of 13 amino  
acids follows the first 7 amino acids. However, the actual  
site of cleavage has not been determined as attempts to  
15 obtain N-terminal sequence of the core protein were hindered  
by a blocked amino terminus. Following the signal sequence  
and preceding the first SmaI site (which is used to define  
the beginning of the tandem repeat region) are 107 amino  
acids. Greater than 50% of these amino acids comprise  
20 degenerate tandem repeats. Since the number of tandem  
repeats per molecule is large (greater than 21 for the  
smallest allele we have observed), this domain forms the  
major part of the core protein, and results in a highly  
repetitive structure which is extremely immunogenic  
25 [Gendler, S. et al., loc. cit]. The sequence of the 20

- 29 -

amino acid tandem repeat unit corresponds to what might be expected for a protein which is extensively O-glycosylated. Five serines and threonines, four of which are in doublets, are found in the repeat and these potential glycosylation  
5 sites are separated by regions rich in prolines (See Fig. 2).

- 30 -

CLAIMS

1. A nucleic acid fragment comprising a portion of at least 17 contiguous nucleotide bases which portion has a sequence the same as, or homologous to a portion of  
5 corresponding length of the sequence of the coding strand as set out in Fig. 1 or the same as, or homologous to a portion of corresponding length of the sequence complementary to the sequence of the coding strand set out in Fig. 1.
2. A fragment according to claim 1 comprising any one or  
10 more of the following:
  - (a) a signal sequence  
TTCCTGCTGCTGCTCCTCACAGTGCTTACAGXTTGTT  
wherein X is an optionally present intron
  - (b) a mammary consensus sequence AGGCTAAACTAGACC
  - 15 (c) a mammary consensus sequence GTAAGAATTGCAGACA
  - (d) a homologue of a sequence (a), (b) or (c) and
  - (e) a sequence complementary to a sequence (a), (b), (c) or (d).
3. A hybridisation probe comprising a fragment according  
20 to claim 1 or claim 2 bearing a detectable label or linked to a solid support.
4. A cloning or expression vector comprising a fragment



- 31 -

according to claim 1 or claim 2.

5. A transformed cell comprising a cloning or expression vector according to claim 4.

6. A polypeptide comprising a sequence of at least 5  
5 contiguous acid residues encoded by the coding portion of the DNA sequence as indicated in Fig. 2.

7. An antibody against a polypeptide according to claim  
6.

8. An antibody according to claim 7 bearing a detectable  
10 label or linked to a solid support.

9. An antibody-producing cell capable of secreting an antibody according to claim 7.

10. A diagnostic kit comprising a fragment according to  
claim 1 or claim 2 or a probe according to claim 3 or a  
15 polypeptide according to claim 6 or an antibody according to claim 7 or claim 8.

11. A fragment according to claim 1 or claim 2 or a probe according to claim 3 or a vector according to claim 4 or a cell according to claim 5 or claim 9 or a polypeptide

- 32 -

according to claim 6 or an antibody according to claim 7 or claim 8 for use in a method of treatment or diagnosis practised on the human or animal body.

12. Use of a fragment according to claim 1 or claim 2 or a probe according to claim 3 or a vector according to claim 4 or a cell according to claim 5 or claim 9 or a polypeptide according to claim 6 or an antibody according to claim 7 or claim 8 in the preparation of a medicament for use in a method of treatment or diagnosis practised on the human or animal body.

13. A method of treatment or diagnosis comprising administering to a cancer patient in need thereof or suspected to have a cancer an effective non-toxic amount of a fragment according to claim 1 or claim 2 or a probe according to claim 3 or a vector according to claim 4 or a cell according to claim 5 or claim 9 or a polypeptide according to claim 6 or an antibody according to claim 7 or claim 8.

14. A method of diagnosis comprising contacting a sample from a patient with a fragment according to claim 1 or claim 2 or a probe according to claim 3 or a vector according to claim 4 or a cell according to claim 5 or claim 9 or a polypeptide according to claim 6 or an antibody according to claim 7 or claim 8.

-803  
 tactctctc cgcccggtcc gagcggcccc tcagttgcg cggcccagcc ccgcaaggct cccggtgacc actagagggc  
 -703  
 gggaggagct cctggccagt ggtggagagt ggaaggaag gaccctaggg ttcatcgga cccaggttta ctcccttaag  
 -603  
 tggaatttc ttcccccaact cctccttggc ttctccaag gagggaaccc aggcctgctgg aaagtccggc tggggggggg  
 -503  
 actgtggtt cagggagaa cggggtgtg aacgggacag ggagcggta gaagggtgg gctattccg gaagtgggtg  
 -453  
 gggaggagg cccaaaacta gcacctagtc cactcattat ccagccctct tatttctcgg ccgctctgct tcagtgagcc //6  
 -403  
 cgggagggc ggggaagtgg agtgggagac ctagggggtgg gcttcccagc ctgctgtac aggacctga cctagctggc  
 -303  
 ttgttcccc atccccacgt tagttgtgc cctgaggcta aaactagagc ccaggggccc caagtccag actgccccctc  
 -203  
 cccctcccc cggagccagg gaglgttgg tgaaggggg aggccagctg gagaacaaac gggtagtcag ggggttgage  
 -153  
 gattagacc ctgtaccct acccaggaat. ggttggggag gaggaggaag aggtaggagg taggggaggg ggcgggggtt  
 -53  
 tgtcacctgt cacctgctcg ctgtgcctag ggcggggcgg cggggagtg ggggaccggt ataaagcgg aggcgcctgt  
 -1+1  
 gccggctcca cctctcaagc agccagggcc tgcctgaatc tgttctgccc cctccccacc catttaccca ccaccatgac  
 +57

Fig.1

SUBSTITUTE SHEET

2/6

```

107      ACCGGGCACC CAGTCTCCIT TCTTCCIGCT GCTGCTCCTC ACAGTGTCTA CAGgtgagg gcacgaggtg gggagtgggc 157
      tgccctgctt aggtggtctt cgtggtcttt ctgtgggttt tgctccctgg cagatggcac catgaagtta aggtaagaat
      +257
      tgcagacaga ggctgcccctg tctgtgccag aaggaggagg aggctaagga caggctgaga agagtggccc ccaaccctga
      +307
      gagtgggtac caggggcaag caaatgtcct gtagagaagt ctagggggaa gagagttagg agagggaagg cttaagaggg
      +407
      gaagaaatgc agggggccatg agccaaggcc tatgggcaga gagaaggagg ctgctgcagg gaaggaggct tccaaccacg
      +507
      gggttactga ggctgcccac tccccagtc tcttggtatt atttctctgg tggccagagc ttatatattc ttcttgctct
      +607
      tatttttctt tcataaagac ccaaccctat gactttaact tcttacagct accacagccc ctaaaccgcg aacagTTGTT
      +657
      ACAGGTTCTG GTCATGCAAG CTCTACCCCA GGTGGAGAAA AGGAGACTTC GGCTACCCAG AGAAGTTCAG TGCCACAGCTC
      +757
      TACTGAGAAG AATGCTGTGA GTATGACCAG CAGCGTACTC TCCAGCCACA GCCCCGGTTC AGGCTCCTCC ACCACTCAGG
      +807
      GACAGGATGT CACTCTGGCC CCGGCCACGG AACCAGCTTC AGGTTACGCT GCCACCTGGG GACAGGATGT CACCTCGGTC
      +907
      CCAGTCACCA GGCCAGCCCT GGGCTCCACC ACCCGGCCAG CCCAGGATGT CACCTCAGCC CCGGACAACA AGCCAGCCCC
      +957
      GGG
      +1007

```

*Fig.1 Cont'd*

	CCG CTC CAC CTC TCA AGA GCC AGC GCC TGC CTG AAT	36
CTG TTC TGC CCC CTC CCC ACC CAT TTC ACC ACC ACC ATG ACA CCG GGC ACC CAG TCT CCT		96
	T P G T Q S P	
	SIGNAL SEQUENCE	
TTC TTC CTG CTG CTC CTC ACA GTG CTT ACA GTT GTT ACA GGT TCT GGT CAT GCA AGC		156
F F L L L L L L T V L T V V T G S G H A S		
	3/6	
TCT ACC CCA GGT GGA GAA AAG GAG ACT TCG GCT ACC CAG AGA AGT TCA GTG CCC AGC TCT		216
S T P G G E K E T S A T Q R S S V P S S		
ACT GAG AAG AAT GCT GTG AGT ATG ACC AGC AGC GTA CTC RCC AGC CAC AGC CCC GGT TCA		276
T E K N A V S M T S S V L S S H S P G S		
GGC TCC TCC ACC ACT CAG GGA CAG GAT GTC ACT CTG GCC CCG GCC ACG GAA CCA GCT TCA		336
G S T T Q G Q D V T L A P A T E P A S		
GGT TCA GCT GCC ACC TGG GGA CAG GAT GTC ACC TCG GTC CCA GTC ACC AGG CCA GCC CTG		396
G S A A T W G Q D V T S V P V T R P A L		
GGC TCC ACC ACC CCG CCA GCC CAG GAT GTC ACC TCA GCC CCG GAC AAC AAG CCA GCC CCG		456
G S T T P P A H D V T S A P D N K P A P		
	SmaI	
	GGC	
	G	

Fig. 2

4/6

```

-608      TTGCTTCTCC  AAGAAGGGAA  CCCAGGTCCG  TGAAGTCCG  GCTGGCGGG  ACTGTGGTT  TACGGGTAGA  -538
-537      ACTCGGIGTG  GAACCGAACG  GGAGCGGTTA  GAAGGGTGGG  GCTATTCGG  AAGTCGTGG  GGGAGGGAGC  -468
-467      CCAAAACTAG  CACCTAGTCC  ACTCATTATC  CAACCGTCTT  ATTCTCGGC  CGCTCTGCTT  CAGTGGACCC  -398
-388      GGGGAGGGGC  GGGGAAGTGG  AGTGGGAGAC  CTAGGGGTGG  GCTTCCCGAC  CTTGCTGTAC  AGGACCTCGA  -328
-327      CCTAGCTGGC  TTTGTTCGCC  ATCCCCACGT  TAGTTGTTGC  CCTGAGGCTA  AAACCTAGACC  GCAGGGGCCC  -258
-257      CAAGTTCCAG  ACTGCCCTCC  CCTCCCCTGG  AGCCAGGGAG  TGGTTGGTGA  AAGGGGAGGC  CAGCTGGAGA  -188
-187      ACAAAACGGT  AGTCAGGGGG  TTGAGCAGTT  AGAGCCCTTG  TACCCTACCC  AGGAATGGTT  GGGAGGAGGA  -118
-117      GGAAGAGGTA  GGAGGTAGGG  GAGGGGCGGG  GGTTTTGTC  CCTGTCACT  GCTCGCTGTG  CCTAGGGCGG  -48

```

Fig.3

SUBSTITUTE SHEET

5/6

47  
 GCGGGCGGGG AGTGGGGGA CCGGTATAAA GCGGTAGCG CCGTGGC 1  
 CCGCTC CAC CTC TCA ACA GCC  
 AGC GCC TGC CTG AAT CTG TTC TGC CCC CTC CCC ACC CAT TTC ACC ACC ACC ATG ACA CCG GGC ACC CAG 73  
 97 SIGNAL SEQUENCE 131 INTRON 1  
 TCT CCI TTC TTC CTG CTG CTG CTC CTC ACA GTG CTT ACA G GT GAGGGGCAC  
 GAGGTGGG AGTGGGCTT GCCCTTGCCT AGGTGGTCT TCGTTGGTTC TTCTGTGGG CTTTGTCTCC  
 CTGGCAGATG GCACCAATGAA GTTAAGGTA GAATATCAGA CAGAGGCTGC CCTGTCTGTG CCAGAAGGAG  
 GGAGAGGCTA AGGACAGGCT GAGAAGAGTT GCGCCCAACC CTGAGAGTGG GTACCAGGG CAAGCAAATG  
 TCCTGTAGAG AAGTCTAGG GGAAGAGAGT AGGGAGAGGG AAGGCTTAAG AGGGAAGAA ATGCAGGGGC  
 CATGAGCCAA GGCCTATGG CAGAGAGAAG GAGGCTGCTG CAGGAAGGA GGCTTCCAAC CCAGGGGTTA

Fig.3 Cont'd(1)

CTGAGGCTGC CCACTCCCCA GTCCICCTGG TATTATTCT CTGGTGGCA GAGCTTATAT TTTCTTCTTG  
CTCTTATTTT TCCTTCAIAA AGACCCAACC CTAIGACTTT AACITCTTAC AGCTACCACA GCCCCTAAAC

641 EXON 2

640

CCGCAACAG TT GTT ACA GGT TCT GGT CAI GCA AGC TCT ACC CCA GGT GGA GAA AAG GAG

6/6

ACT IOG GCT ACC CAG AGA AGT TCA GTG CCC AGC TCT ACT GAG AAG AAT GCT GTG AGT ATG ACC AGC AGC

GTA CIC TCC AGC CAC AGC CCC GGT TCA GGC TCC CCA CCA CIC AGG GAC AGG ATG TCA CIC IGG CCC CGG

CCA CGG AAC CAG CTT CAG GTT CAG CIG CCA CCT GGG GAC AGG ATG TCA CCT CGG TCC CAG TCA CCA GGA

GCC CTG GGC TCC ACC ACC GGG CCA GCC CAC GAT GTG ACC TCA GCC CCG GAC AAC AAG CCA GCC CCG GG

968

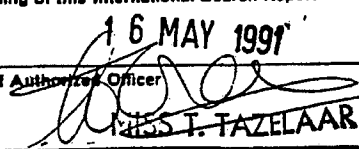
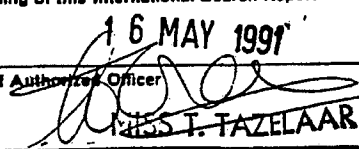
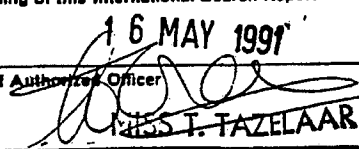
-----> TANDEM REPEAT

Fig.3 Cont'd(2)



# INTERNATIONAL SEARCH REPORT

International Application No PCT/GE 90/02020

<b>I. CLASSIFICATION OF SUBJECT MATTER</b> (If several classification symbols apply, indicate all) <sup>6</sup> According to International Patent Classification (IPC) or to both National Classification and IPC IPC <sup>5</sup> : C 07 H 21/04, C 07 K 13/00, C 12 N 15/12, C 12 P 21/00, C 12 Q 1/68, G 01 N 33/574, C 07 K 15/00																	
<b>II. FIELDS SEARCHED</b> <div style="text-align: center;">Minimum Documentation Searched <sup>7</sup></div> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <th style="width: 25%;">Classification System</th> <th>Classification Symbols</th> </tr> <tr> <td style="text-align: center; vertical-align: top;">IPC<sup>5</sup></td> <td style="text-align: center; vertical-align: top;">C 07 K, C 12 N, C 12 P, C 12 Q</td> </tr> </table> <p style="text-align: center; font-size: small;">Documentation Searched other than Minimum Documentation to the extent that such Documents are included in the Fields Searched <sup>8</sup></p>			Classification System	Classification Symbols	IPC <sup>5</sup>	C 07 K, C 12 N, C 12 P, C 12 Q											
Classification System	Classification Symbols																
IPC <sup>5</sup>	C 07 K, C 12 N, C 12 P, C 12 Q																
<b>III. DOCUMENTS CONSIDERED TO BE RELEVANT<sup>9</sup></b> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <th style="width: 10%;">Category <sup>9</sup></th> <th style="width: 60%;">Citation of Document, <sup>11</sup> with indication, where appropriate, of the relevant passages <sup>12</sup></th> <th style="width: 30%;">Relevant to Claim No. <sup>13</sup></th> </tr> <tr> <td style="text-align: center; vertical-align: top;">X</td> <td style="vertical-align: top;">WO, A, 8805054 (IMPERIAL CANCER RESEARCH TECHNOLOGY) 14 July 1988 see the whole document; especially claims cited in the application</td> <td style="text-align: center; vertical-align: top;">6</td> </tr> <tr> <td style="text-align: center; vertical-align: top;">Y</td> <td style="text-align: center; vertical-align: top;">--</td> <td style="text-align: center; vertical-align: top;">3-5,7-10,14</td> </tr> <tr> <td style="text-align: center; vertical-align: top;">P,X</td> <td style="vertical-align: top;">Journal of Biological Chemistry, vol. 265, no. 10, 5 April 1990, The American Society for Biochemistry and Molecular Biology, Inc., (US), M.J.L. Ligtenberg et al.: "Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene encoding splice variants with alternative amino termini", pages 5573- 5578 see the whole article</td> <td style="text-align: center; vertical-align: top;">1,2,6-9</td> </tr> <tr> <td style="text-align: center; vertical-align: top;">P,Y</td> <td style="text-align: center; vertical-align: top;">--</td> <td style="text-align: center; vertical-align: top;">3-5,10,14</td> </tr> </table>			Category <sup>9</sup>	Citation of Document, <sup>11</sup> with indication, where appropriate, of the relevant passages <sup>12</sup>	Relevant to Claim No. <sup>13</sup>	X	WO, A, 8805054 (IMPERIAL CANCER RESEARCH TECHNOLOGY) 14 July 1988 see the whole document; especially claims cited in the application	6	Y	--	3-5,7-10,14	P,X	Journal of Biological Chemistry, vol. 265, no. 10, 5 April 1990, The American Society for Biochemistry and Molecular Biology, Inc., (US), M.J.L. Ligtenberg et al.: "Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene encoding splice variants with alternative amino termini", pages 5573- 5578 see the whole article	1,2,6-9	P,Y	--	3-5,10,14
Category <sup>9</sup>	Citation of Document, <sup>11</sup> with indication, where appropriate, of the relevant passages <sup>12</sup>	Relevant to Claim No. <sup>13</sup>															
X	WO, A, 8805054 (IMPERIAL CANCER RESEARCH TECHNOLOGY) 14 July 1988 see the whole document; especially claims cited in the application	6															
Y	--	3-5,7-10,14															
P,X	Journal of Biological Chemistry, vol. 265, no. 10, 5 April 1990, The American Society for Biochemistry and Molecular Biology, Inc., (US), M.J.L. Ligtenberg et al.: "Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene encoding splice variants with alternative amino termini", pages 5573- 5578 see the whole article	1,2,6-9															
P,Y	--	3-5,10,14															
<div style="display: flex; justify-content: space-between;"> <div style="width: 45%;"> <p><sup>10</sup> Special categories of cited documents:</p> <p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier document but published on or after the international filing date</p> <p>"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p> </div> <div style="width: 45%;"> <p>"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step</p> <p>"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.</p> <p>"Δ" document member of the same patent family</p> </div> </div>																	
<b>IV. CERTIFICATION</b> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; padding: 5px;">           Date of the Actual Completion of the International Search  <div style="text-align: center;">3rd April 1991</div> </td> <td style="width: 50%; padding: 5px;">           Date of Mailing of this International Search Report  <div style="text-align: center;">16 MAY 1991</div> </td> </tr> <tr> <td style="width: 50%; padding: 5px;">           International Searching Authority  <div style="text-align: center;">EUROPEAN PATENT OFFICE</div> </td> <td style="width: 50%; padding: 5px;">           Signature of Authorized Officer  <div style="text-align: center;">               J. TAZELAAR           </div> </td> </tr> </table>			Date of the Actual Completion of the International Search <div style="text-align: center;">3rd April 1991</div>	Date of Mailing of this International Search Report <div style="text-align: center;">16 MAY 1991</div>	International Searching Authority <div style="text-align: center;">EUROPEAN PATENT OFFICE</div>	Signature of Authorized Officer <div style="text-align: center;">               J. TAZELAAR           </div>											
Date of the Actual Completion of the International Search <div style="text-align: center;">3rd April 1991</div>	Date of Mailing of this International Search Report <div style="text-align: center;">16 MAY 1991</div>																
International Searching Authority <div style="text-align: center;">EUROPEAN PATENT OFFICE</div>	Signature of Authorized Officer <div style="text-align: center;">               J. TAZELAAR           </div>																

III. DOCUMENTS CONSIDERED TO BE RELEVANT (CONTINUED FROM THE SECOND SHEET)		
Category *	Citation of Document, " with indication, where appropriate, of the relevant passages	Relevant to Claim No.
P,X	Journal of Biological Chemistry, vol. 265, no. 25, 5 September 1990, The American Society for Biochemistry and Molecular Biology, Inc., (US), S.J. Gendler et al.: "Molecular cloning and expression of human tumor- associated polymorphic epithelial mucin", pages 15286-15293 see figure 1 --	1,2,6
P,X	Journal of Biological Chemistry, vol. 265, no. 25, 5 September 1990, The American Society for Biochemistry and Molecular Biology, Inc., (US), M.S. Lan et al.: "Cloning and sequencing of a human pancreatic tumor mucin cDNA", pages 15294-15299 see the whole article --	1,2,6-9
P,A	WO, A, 9005142 (IMPERIAL CANCER RESEARCH TECHNOLOGY) 17 May 1990 see abstract and claims cited in the application -----	1-10,14

## FURTHER INFORMATION CONTINUED FROM THE SECOND SHEET

V. ☒ OBSERVATIONS WHERE CERTAIN CLAIMS WERE FOUND UNSEARCHABLE <sup>1</sup>

This International search report has not been established in respect of certain claims under Article 17(2) (a) for the following reasons:

1. ☒ Claim numbers XX..... because they relate to subject matter not required to be searched by this Authority, namely:

Claims 11-13

Pls. see Rule 39.1 (iv) - PCT:

Methods for treatment of the human or animal body by surgery or therapy, as well as diagnostic methods.

2. ☐ Claim numbers ....., because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claim numbers ....., because they are dependent claims and are not drafted in accordance with the second and third sentences of PCT Rule 6.4(a).

VI. ☐ OBSERVATIONS WHERE UNITY OF INVENTION IS LACKING <sup>2</sup>

This International Searching Authority found multiple inventions in this international application as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims of the international application.
2. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims of the international application for which fees were paid, specifically claims:
3. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claim numbers:
4. ☐ As all searchable claims could be searched without effort justifying an additional fee, the International Searching Authority did not invite payment of any additional fee.

## Remark on Protest

- ☐ The additional search fees were accompanied by applicant's protest.
- ☐ No protest accompanied the payment of additional search fees.

**ANNEX TO THE INTERNATIONAL SEARCH REPORT  
ON INTERNATIONAL PATENT APPLICATION NO.**

GB 9002020  
SA 43255

This annex lists the patent family members relating to the patent documents cited in the above-mentioned international search report. The members are as contained in the European Patent Office EDP file on 07/05/91  
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO-A- 8805054	14-07-88	AU-A- 1103988	27-07-88
		EP-A- 0341252	15-11-89
		JP-T- 2501828	21-06-90
-----			
WO-A- 9005142	17-05-90	None	
-----			